DETAILED MATERIALS AND METHODS

Description and genotyping of patient cohort

Patients included in this study were diagnosed with CD and underwent ileo-colic resection at Washington University School of Medicine between 2005 and 2010 or at Cedars-Sinai Medical Center between 1999 and 2010. De-identified tissue samples from the proximal resection margins, genomic DNA isolated from a blood sample and the demographic information associated with the cases were obtained from the Washington University Digestive Diseases Research Core Center (DDRCC) Tissue Procurement Facility, or from the IBD Center at Cedars-Sinai Medical Center. Inclusion criteria were that the proximal margin tissue sample had 1) at least 100 well-oriented intestinal crypts (not necessarily contiguous) and 2) absent or minimal active or chronic inflammatory disease (cryptitis, crypt abscess, architectural distortion, pyloric gland metaplasia, etc.) as determined by pathologists (T.S.S. and T.C.L.). Regions of well-oriented crypts were defined as those that contained crypt lumens that extended from the apex to the base of the crypt. Patient DNA samples were genotyped for *ATG16L1 T300A* and the CD-associated *NOD2* variants[1, 2]. The *NOD2* variants included both common (*R702W*, *G908R* and *L1007fsXinsC*) and rare (*R311W*, *S431L*, *R703C*, *V793M*, *N852S* and *M863V*) CD-associated variants[3]. Patients from the Washington University cohort were genotyped by the Washington University DDRCC using matrix-assisted laser desorption ionization-time of flight mass spectrometry and by the Genome Technology Access Center using the Human OmniQuad SNP genotyping arrays (Illumina). Patients from the Cedars-Sinai cohort were genotyped using the Immunochip (Illumina). Medical records were reviewed to identify the time to disease recurrence post-resection surgery for CD cases 1-178 (Tables S1, S2 and S3). Disease recurrence was defined by endoscopy[4] and/or histology. Patients who received immunomodulators (e.g., 6-mercaptopurine, azathioprine, methotrexate) or biologics (i.e., anti-TNF therapy) post-resection were considered to have received prophylaxis. Paneth cell morphological scoring data, genotype data, granuloma data and disease recurrence data for each patient are listed in Tables S1, S2 and S3. Patients 1-59 (*ATG16L1 T300A* 'safe' cohort) are listed in Table S1; these cases were used to generate the data in Figures 1-5. Patients 60-119 (ATG16L1 T300A 'risk' cohort) are listed in Table S2; these cases were used to generate the data in Figures 2-5. Patients 120-178 (non-genotyped cohort) are listed in Table S3; these cases were used to generate the data in Figure 5. The study was approved by the Institutional Review Boards of Washington University School of Medicine and Cedars-Sinai Medical Center. Written informed consent was obtained from all study participants.

Morphological analysis of Paneth cells

Formalin-fixed paraffin-embedded (FFPE) tissue sections were immunostained with primary antibody lysozyme C-19 (Santa Cruz sc-27958). Lysozyme distribution was quantified as previously described[5]. For each case, a pathologist who was blinded to the identity of the cases (T.C.L.) scored a minimum of 200 Paneth cells (range: 206-2,702) in well-oriented crypts. Paneth cells were categorized as normal or abnormal (disordered, diminished, diffuse or excluded granule). Paneth cells were scored as normal if they contained numerous small (~1 μm), lysozyme-positive granules located in the apical half of the cell. Paneth cells were scored as abnormal if the lysozyme staining pattern was disordered, diminished or diffuse. Disordered Paneth cells contained lysozyme-positive granules that were of normal size and quantity, but had some basally located granules. Diminished Paneth cells contained <10 granules, and the remaining granules were frequently enlarged or fused. Diffuse Paneth cells did not contain any secretory granules, but rather showed diffuse lysozyme staining throughout their cytoplasm.

Excluded granule Paneth cells contained granule shapes that had low/absent lysozyme staining and diffuse cytoplasmic lysozyme staining with occasional lysozyme-positive granules. Paneth cells located within Peyer's patches were excluded. Paneth cells were scored with an Olympus BX41 microscope equipped with a DP72 digital camera using the 60X objective lens. Microscopy was performed with a Zeiss Axiovert 200M inverted microscope equipped with an Axiocam MRm digital camera.

Microarray Analysis

FFPE tissue sections (5 µM) were prepared from the set of archived surgical resection samples used for histological analysis in this study. Whole ileal material was procured by scraping two tissue sections per sample (n = 40). Laser capture microdissection (LCM) was performed using the Arcturus system and CapSure HS LCM caps (Applied Biosystems); methyl green stain was used to identify Paneth cells (~8,000 Paneth cells were microdissected per sample; n = 15). For both sample types, RNA extraction was performed using the RNeasy FFPE kit (Qiagen) according to the manufacturer's instructions and 30 µL of ddH_20 was used for elution. For cDNA synthesis and amplification with the Transplex Whole Transcriptome Amplification kit (WTA2; Sigma), each RNA sample was split equally between two reactions. Subsequent purification of the cDNA product was performed with the PCR Purification kit (Qiagen). For this step, the two WTA2 reactions per sample were combined and run through a single purification column. Cy5 labeling with the ULST Fluorescent Labeling kit (Kreatech) and hybridization to Whole Human Genome 4x44k Microarrays (Agilent) was performed by the Genome Technology Access Center at Washington University – St. Louis.

To determine the set of genes enriched in Paneth cells of CD patients, the set of transcriptional profiles generated from laser capture microdissected Paneth cells was compared to the set of transcriptional profiles generated from whole ileal tissue. To identify transcripts from the whole ileal tissue data sets with significant association to particular Paneth cell phenotypes, Pearson's product moment correlation coefficient (PCC) was used to measure the correlation between the sets of phenotype scores and the normalized transcript expression level for each Paneth cell-enriched gene using the following formula: PCCgeneX,granulePhenotypeY = cov(X,Y)/sd(X)*sd(Y), where X is the vector that contains the expression level of gene X and Y is the vector that contains the corresponding granule phenotype score Y for each of the patients. The significance of each correlation was calculated using a Student's t-distribution for a transformation of the correlation as implemented in MATLAB with P <0.05. Analysis of GO terms was performed using DAVID Bioinformatic Resource (NIAID/NIH)[6, 7]. Data are deposited at ArrayExpress (accession number E-MTAB-1281).

Correlation analyses

For clustering of the CD patients, each case was represented as a numerical vector containing the observations for each of the following parameters: Paneth cell phenotypes, genetics, documented environmental exposures and demographic information. The Pearson correlation between patient vectors was calculated using GENE-E[8]. Heat map cells were generated using unsupervised clustering and colored according to the Pearson correlation, with red and blue cells representing positive and negative correlations, respectively. A marker selection strategy[9] implemented in GENE-E was used to identify clinical variables with an FDR-adjusted $P \leq 0.05$ that were correlated with the two patient subtypes.

For the correlation of Paneth cell phenotypes, genetics, documented environmental exposures and demographic parameters, each variable was represented as a numerical vector of observations and then the Pearson correlation between the vectors was calculated using GENE-E[8]. The resulting correlation matrix was transformed by determining the Z-score for each row. Heat map cells were generated using unsupervised clustering and colored according to the 95% confidence values of the Z-scores, with red representing positively correlated variables and blue representing negatively correlated variables). The *P* values for each Pearson's correlation value were calculated using a Student's t-distribution with $n - 2$ degrees of freedom (where n is the number of patients).

Statistical analyses

Comparison of the demographic parameters between patients without *NOD2* risk variants and those with one or more *NOD2* risk variants was performed using Fisher's exact test. For the analysis of lysozyme quantification, permutation tests were performed to test the association between *NOD2* variants and the percentage of abnormal Paneth cells. In short, the phenotypes were randomly permuted 1,000 times across all the subjects and a linear regression was performed for each permutation to create a null distribution of the test statistics. An adjusted p-value was then calculated by comparing this distribution to the statistics without permutation. Mann-Whitney tests were used to demonstrate statistical difference between cases with 1 or 2 NOD2 risk variants and controls. Linear regression was used to analyze the cumulative number of risk variants. A Chi-Square test and a log-rank test were performed for the analysis of granuloma incidence and time to disease recurrence, respectively, with $P < 0.05$ considered to be significant (Prism GraphPad software).

REFERENCES

1. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118-25.
2. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119-24.
3. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 2011;43:1066-73.
4. Daperno M, D'Haens G, Van Assche G, et al. Development and validation of a new, simplified endoscopic activity score for Crohn's disease: the SES-CD. Gastrointest Endosc 2004;60:505-12.
5. Cadwell K, Liu JY, Brown SL, et al. A key role for autophagy and the autophagy gene Atg16l1 in mouse and human intestinal Paneth cells. Nature 2008;456:259-63.
6. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 2009;37:1-13.
7. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 2009;4:44-57.
8. Gehlenborg N, Wong B. Heat maps. Nature Methods 2012;9:213-213.
9. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98:15149-54.